DOCUMENT RESUME

ED 150 169 TM 006 872

AUTHOR	Tallmadge, G. K.; Horst, D. P.
TITLE	Using the Data from State and Local ESEA Title I Reports.
PUB DATE	Apr 77
NOTE	16p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977) ; Best copy available
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS	*Academic Achievement; Achievement Tests; *Compensatory Education Programs; *Educationally Disadvantaged; Elementary Education; Federal Aid; Grade Equivalent Scores; *Information Utilization; Mathematics; Norm Referenced Tests; *Program Evaluation; Reading; Research Problems; School Districts; Standardized Tests; State Programs; Testing; *Testing Problems
IDENTIFIERS	*Elementary Secondary Education Act Title I

ABSTRACT

Many of the evaluations summarized in state and local Elementary and Secondary Education Act (ESEA) Title I reports have been negatively affected by methodological errors or the use of different, incompatible formats. Although new, detailed evaluation and reporting procedures have been developed for Title I research projects, two problems remain: bias caused by the tendency to report only favorable results, and inadequacies in the instruments used to measure achievement gain. Some suppression of negative evaluation findings will probably always exist, motivated either by fear of the repercussions, face saving, or the sincere belief that something is wrong with the data. Therefore, it is likely that all aggregations of evaluation data will produce systematically inflated results. It may, however, be possible to estimate the extent of the bias. The testing methods also need improvement. Standardized tests can be chosen to be as relevant as possible to what the students are being taught; and criterion- or objectives-referenced tests may offer a solution to the insensitivity problem--but at the expense of norm-referencing. In short, there is no effective way to interpret large, norm-referenced gains in isolated skill areas. (Author/MV)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED150169

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Kasten Tallmadge
RMC Research Corp.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

USING THE DATA FROM STATE AND LOCAL ESEA TITLE I REPORTS

BEST COPY AVAILABLE

G. K. Tallmadge
D. P. Horst

RMC Research Corporation
Mountain View, California

Paper presented at the Annual Meeting of the American Educational Research Association, April 4-8, 1977, New York City.

TM006 872

USING THE DATA FROM STATE AND LOCAL ESEA TITLE I REPORTS

G. K. Tallmadge and D. P. Horst
RMC Research Corporation

Introduction

The title of this paper is somewhat misleading--actually we believe that most of the data currently contained in State or local ESEA Title I reports are far less useful than might be desired. Our concern with the usability of the data reflects the basic differences among the goals and approaches of the three studies described in this symposium. In the case of ETS, data collection was a major part of their study, and they were not involved in analyzing existing data. On the other hand, SRI was charged with learning as much as possible from existing data about Title I impact, while the RMC review of State and local Title I reports was primarily a first step in the development of a more satisfactory Title I evaluation and reporting system. Our main concern was to identify the problems with existing reports. Aggregating and "using" the data contained in those reports was clearly a second priority--although we did study some 250 State reports and many more local ones.

Of course the ultimate goal of all the studies was to provide educational policy makers in Washington with clear and credible answers to such questions as the following:

- Are the students served by Title I any better off than they would have been without the program?
- Are benefits related to the amount or type of service provided?
- Are costs related to gains?
- Is the program more effective at some grade levels than at others?

Many more questions can be, and have been, posed. Of interest, however, is the fact that most of the questions are directly or indirectly concerned with the benefits which may accrue to children as a result of their participation in Title I projects and programs. Furthermore, because most Title I projects focus on compensatory instruction in the basic skill areas of reading and math, the concern with impact can often be expressed as the simple question, "Are kids better able to read and/or do math after participating in a Title I project than they would have been without such participation?" Intuitively, it appears that this question ought to be simple and straightforward to answer, but experience tells us that it is not. While there are many technical problems that stand in the way of finding satisfying answers, we now believe that these technical problems are not the real source of the frustration we have all felt. The real problems have come from inconsistant frames of reference and unrealistic expectations.

Perspective on Cognitive Achievement Gains

Probably all Title I evaluators are well aware of the tremendous individual differences which exist among children at any given grade level. The difference in reading between the 10th- and 90th-percentile students, for example, is readily apparent from the first grade. By the sixth grade, it is the difference between reading well and barely reading at all.

Most of us probably know, or would be willing to assume, that a 10th-percentile sixth grader cannot read nearly as well as a 90th-percentile fifth grader. On the other hand, few of us would know exactly where a 10th-percentile sixth grader would stand with respect to the distribution of fifth graders, yet status indicators of this type provide an extremely meaningful and useful frame of reference for measuring growth.

Figure 1 presents approximate growth curves in reading for grades 2 through 8 for three different students--one who remains at the 50th percentile throughout the seven-year period, one who remains one standard

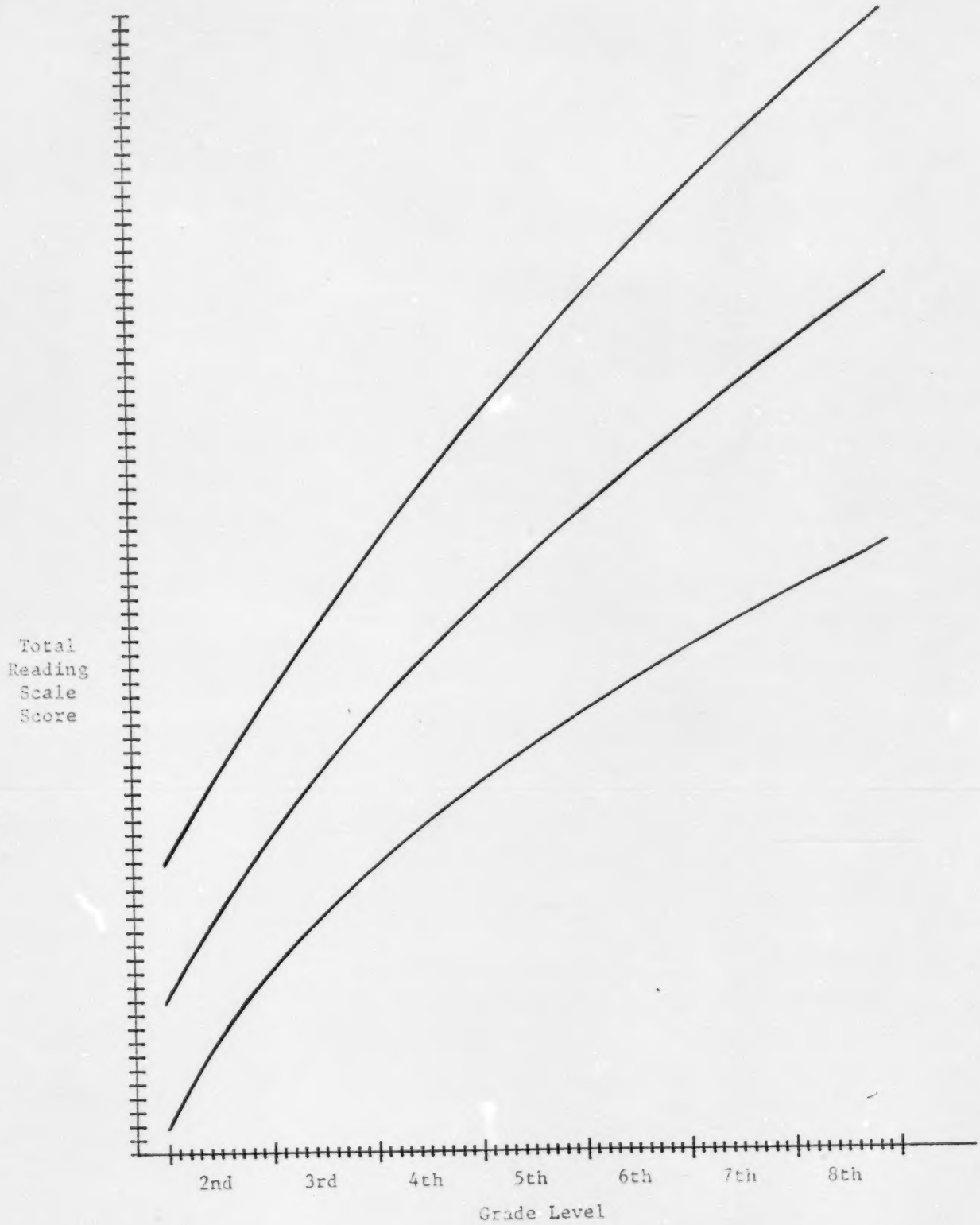

Figure 1. Reading achievement growth curves for 16th, 50th, and 84th percentile students.

deviation above the median, and one who remains one standard deviation below the median. The curves were derived by "tracking" the 16th, 50th, and 84th percentiles through the norms tables of eight widely used, nationally normed achievement tests. Raw and/or standard score scales used by the individual publishers were converted to z-scores to enable aggregation, and the curves have been smoothed for the sake of clarity in this presentation.

There are several interesting things to note about this graph. First, the curves are negatively accelerated, showing that the learning process is quite rapid at the early grades but that it slows down considerably in the later grades. This makes sense. Most of us probably read about as well when we finished high school as we do now. The other thing to note is that the three learning curves get farther apart with increasing grade levels. In fact, they are almost exactly twice as far apart at the end of eighth grade as at the beginning of second. Since the lines are a standard deviation apart, this means that the standard deviation is increasing with time. That, in turn, means that individual differences among children are greater in the later grades than in the earlier grades.

It is possible to express the growth which occurs during the course of any given school year in terms of the standard deviation among students at the end of the year--and it is quite interesting to do so. Figure 2, which shows just a small portion of the growth curves, illustrates these relationships.

In Figure 2, the solid line illustrates the normal growth expectation for an average student in an average school over the course of one year. The dotted horizontal lines are projections of his pretest and posttest scores. The distance between the two lines reflects the number of points gained between pre- and posttest. Dividing this distance by the standard deviation of the population of end-of-year students gives us a standardized gain score.

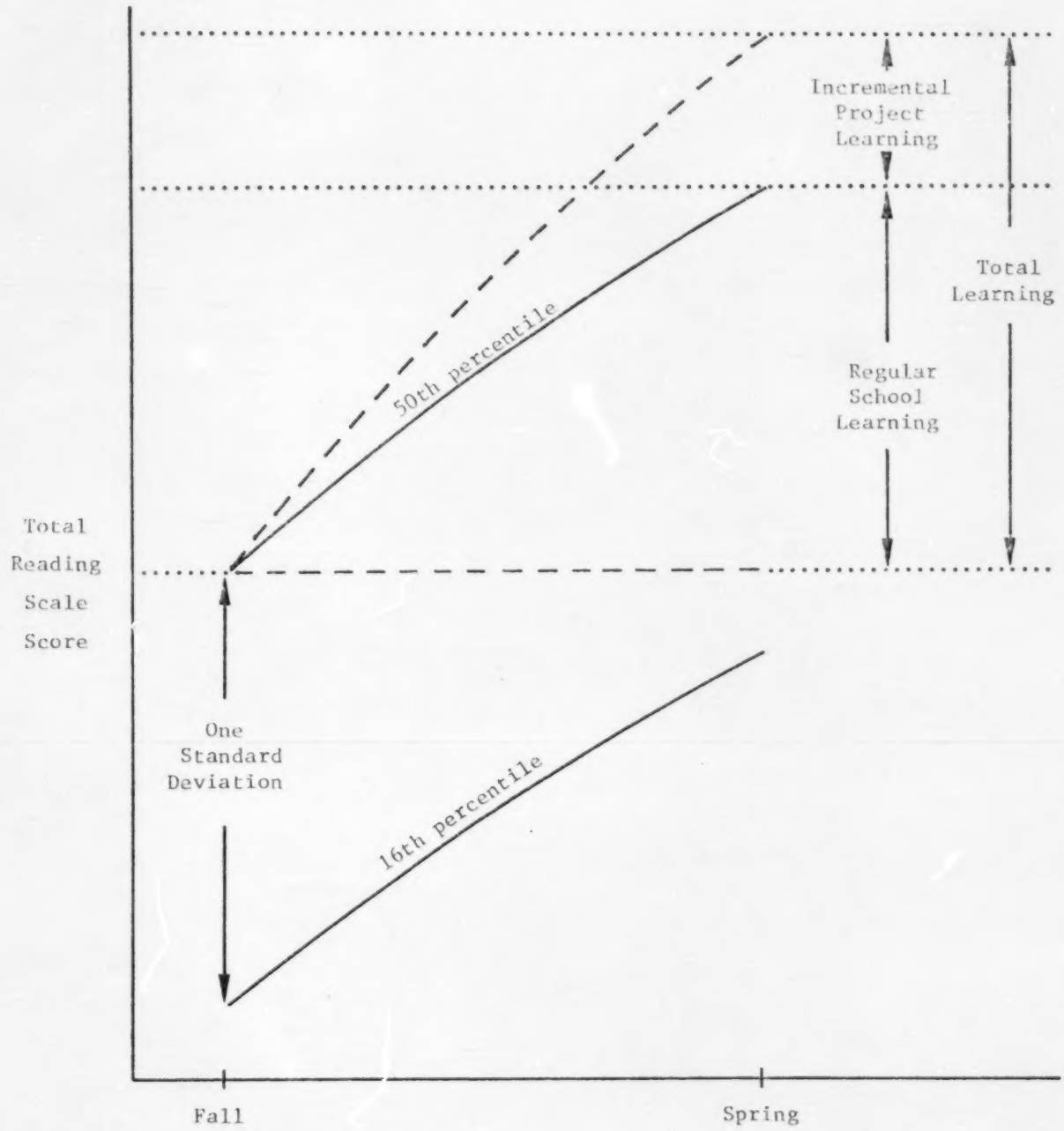

Figure 2. Growth curve for 16th and 50th percentiles during one academic year.

The broken line superimposed over the pretest score projection may be considered as a zero-growth learning curve. The other broken line represents the kind of learning curve which might be observed where an effective special instructional project was supplementing the regular school program--a situation analogous to Title I. If the distance between the zero growth learning curve and the solid line learning curve defines the amount of learning which can be attributed to the regular school project plus maturation and all out-of-school learning, then the distance between the solid line and the supplementary project line defines the increment of learning which can be attributed to the supplementary project. Total learning is simply the sum of the two components.

The relationships among these three learning curves change as a function of grade level, the status of the child with respect to the national norm group, and the effectiveness of both the regular school program and the supplementary project. To simplify the remainder of this discussion, however, we will assume that effectiveness of the regular school program is exactly average and that its impact will be to keep each child exactly on his or her own percentile "track."

Figure 3 superimposes the Figure 2 relationships on the learning curves shown in Figure 1. Zero growth and special project learning curves have been added to the original three regular-school learning curves for the second, fourth, and sixth grades. The time span covered in each of these grades is from October to May--the normal testing interval for fall-to-spring evaluations. In all cases, the special project learning curve has been drawn to represent incremental learning of one-third of a standard deviation--or one-third of the distance between the 50th percentile learning curve and either of the other two. (We chose the one-third standard deviation learning increment somewhat arbitrarily, but it is roughly comparable to the impact of the most effective compensatory education projects we have studied--regardless of the grade levels they served.)

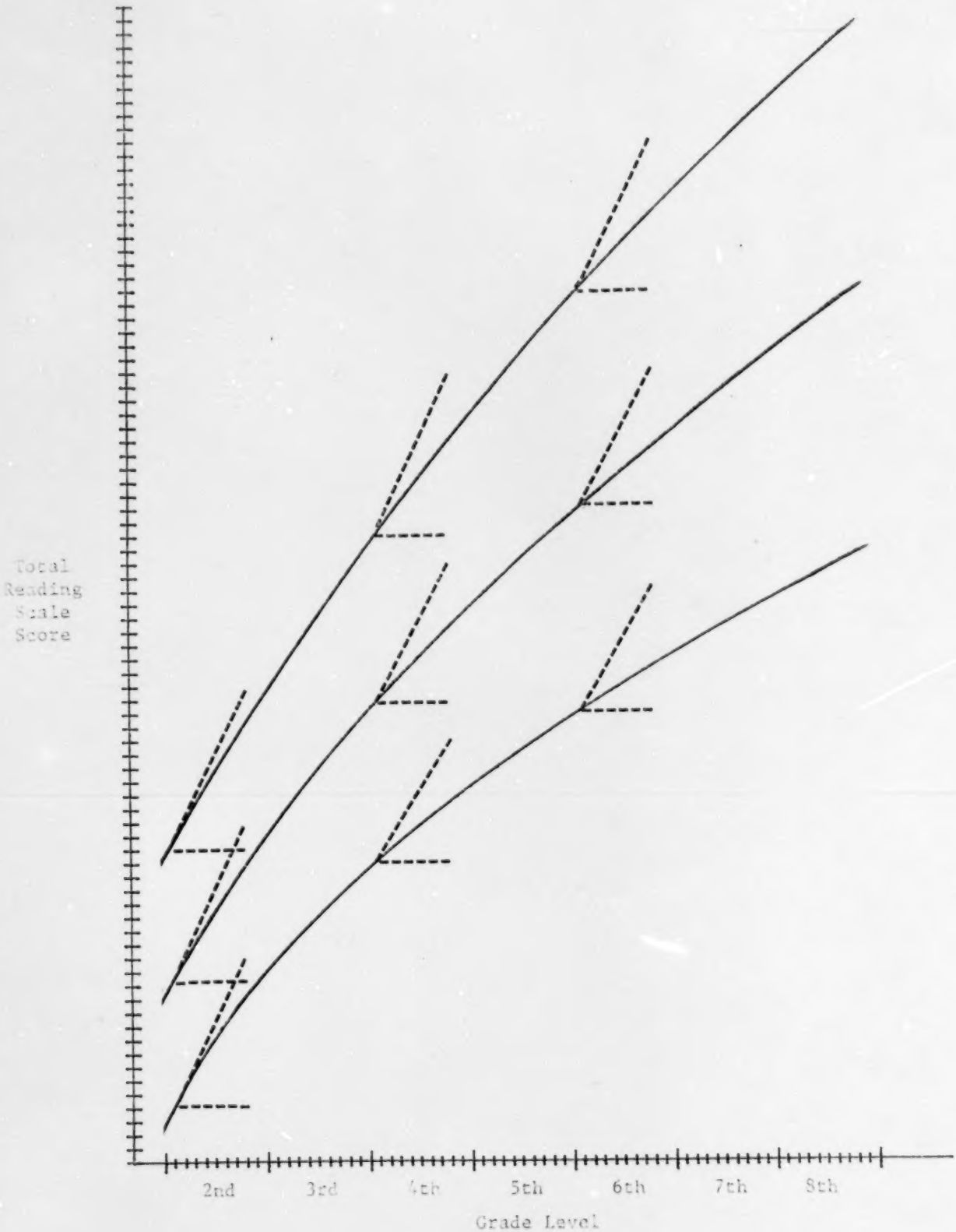

Figure 3. Reading achievement growth curves for 16th, 50th, and 84th percentile students showing no learning and supplementary project expectations in 2nd, 4th, and 6th grades.

Across all grade levels, the regular school gains made by our low-achieving student are smaller than those made by our average student--which in turn are smaller than those made by our high-achieving student. In second grade, however, these differences are not large, and all three students make roughly a whole standard deviation gain. In other words, the difference between their pretest and posttest scale scores is approximately equal to the posttest standard deviation of the normative sample.

In fourth grade, the average student makes a regular school gain of about four-tenths of a standard deviation, while in sixth grade, the gain is about one-quarter of a standard deviation. This means that, at the end of fourth grade, the distance between the 50th and the 84th percentile student is between two and three times as great as the total gain made by an average pupil during the entire fourth-grade year. In sixth grade, the distance between the 50th and the 84th percentiles is four times the average annual growth.

It is interesting to consider growth due to exemplary supplementary instructional programs in the same context. In Figure 3, the special project learning curves (really the regular-school-plus-special-project learning curve) have been drawn to add exactly one-third standard deviation to the regular school learning. Learning attributable to the most effective special projects is thus equal to about one-third of the regular school growth in second grade. It is only slightly less than the regular school growth in fourth grade, and about one-and-a-third times larger than regular growth in sixth grade. And, of course, the ratio is even larger for our hypothetical low-achieving student.

The point of all this is simply to make it clear that when we look at sets of pre- and posttest scores, most of the variance will be within sets, representing pre-existing individual differences among students. Except at second grade, a substantially smaller proportion of the score variance will be attributable to treatment effects, both regular and supplementary. If we are interested in just one or the other of those, we

are likely to be dealing with a very small proportion of the total variance, or to put it another way, if the goal of our compensatory education efforts is to eliminate the very large difference between low- and high-scoring students, we will probably be disappointed by the quite small changes which special projects appear to make.

Just one more example. Suppose, for a group of disadvantaged children we were able to design a compensatory education project which doubled the learning rates of participating students. On the face of it, that would sound like a marvelous project. But if our children were 16th percentile sixth graders, it would take about two and three-quarters years of regular plus supplementary sixth-grade instruction to bring them up to sixth-grade level. And then they would still be two and three-quarters years behind the norm for their age level peers.

Title I Evaluation Problems

It should be clear at this point that the kind of treatment effects we will be looking for in Title I projects will be very small--at least in comparison to individual differences among children in elementary school. To detect small differences, we need precise instruments properly employed. Unfortunately, most often we have crude instruments that are misused. Let's consider the misuses first.

Most locally conducted evaluations today employ what we call a norm-referenced model. In other words, they use normative data provided by test publishers as a substitute for a control or comparison group. Unfortunately, few evaluators think about the parallels between control-group and norm-referenced evaluation models. Thus, when implementing a norm-referenced evaluation, they fail to follow the rules they would observe quite scrupulously if they were implementing a control-group study. They do not worry very much about the comparability of children in the treatment and comparison groups, nor does it often occur to them that the two groups should be pre- and posttested at the same points in time. The solution to this testing time problem is obvious--once the problem has

been recognized: simply test at the times the normative data are collected. Local norms can be developed to maximize the comparability of treatment and comparison children.

Other misuses include using a single set of test scores both as a means of identifying the children with the greatest need for the program and as their pretest measure. This widely employed practice, of course, results in the regression effect error and can produce sizeable but entirely spurious apparent gains.

Another major defect, which until very recently was found almost universally, is the use of grade-equivalent scores. The problem is best illustrated by a concrete example. Many standardized tests are normed in the spring, about the eighth month of the school year. Under these circumstances, the conversion from raw scores to grade equivalents may be empirical at grade levels 1.8, 2.8, 3.8, etc. Other conversions, however, must necessarily be derived through some type of interpolation or extrapolation. To the authors' best knowledge, a linear interpolation or extrapolation has always been used.

Unfortunately, cognitive growth is not linear over the ten-month, grade-equivalent year. As a general rule, low-achieving students make less than one-tenth of their average yearly growth over the summer. In some instances, at least, they actually show losses over the summer, getting fewer items correct in the fall than they did the previous spring on the same test. Under these circumstances, the linear projections will reflect systematic biases. The projected fall norms will show children performing farther below grade level than they really are, and fall-to-spring evaluations will thus spuriously inflate apparent gains. The same kinds of problems plague projected percentile norms but, at least where there are empirical normative data points, the entire range of percentiles can be accurately described. This situation never exists with grade-equivalent scores.

A further problem exists with respect to aggregated evaluation results. If there is any tendency to report positive results while "forgetting" to report negative results, a serious bias can result. If the lowest 40% of the data are excluded from the aggregation, the reported mean will be spuriously inflated by approximately two-thirds of a standard deviation (Roberts, 1977). A program which had a negative impact of one-third standard deviation could thus be made to have an apparent positive impact of one-third standard deviation. While the existence of a bias of this magnitude is unlikely, the absence of any bias is perhaps equally unlikely. Most of the common evaluation errors--particularly those associated with norm-referenced evaluations--have the effect of inflating apparent gains. Actions taken consciously or even unconsciously to make projects look good will, of course, bias the data in the same direction.

It was not long ago when Horst, Tallmadge, and Wood (1975) made the statement that "most standardized reading and math tests are sensitive to any significant cognitive growth and should usually prove adequate for assessing the impact of special treatments" [p. 5]. Those authors no longer hold that position and now believe that standardized tests may not be sufficiently sensitive to detect the comparatively small gains projects are likely to effect--at least where small numbers of pupils are tested. The problem now appears to be to measure the less generalized, relatively small, but still useful skills which the better projects appear to be imparting. While some standardized achievement tests may be adequate for assessing the impacts of some special projects, matches between the content of the test and the content of the curriculum can be quite poor. If most of the test relates to material which has not been taught, it is unlikely to pick up treatment effects which, under optimum conditions, could represent only a small proportion of total score variance. The problem is particularly severe at the lower grade levels where there is little learning generalization beyond the specifics of what is taught.

Bianchini (1976) provides an illustration from California that highlights this issue. Statewide testing in 1970 revealed that the median

first-grade pupil fell at the 38th percentile of the national distribution of scores on the Stanford Reading Test. In 1971, a switch was made to the Cooperative Primary Test, and the median California first grader was right at the national median. Subsequent analyses revealed that there was only about 19% congruence between the vocabulary in the Stanford Reading Test and the vocabulary contained in the instructional materials provided by the State at that time. The congruence was about 55% with the Cooperative Primary Test.

Even 55% congruence seems low. If, indeed, nearly half of the items on a test are irrelevant to the curriculum, then that instrument becomes a very unreliable yardstick for measuring progress in the intended direction.

Conclusions and Recommendations

In looking at the evaluations summarized in State and local ESEA Title I reports, it was clear that many, if not most, of them were contaminated by methodological errors of varying degrees of seriousness. These problems with the data were further compounded by the fact that many different and non-compatible reporting formats were used, thereby complicating or obviating aggregation. To combat these deficiencies and make future reports more useful, step-by-step evaluation and reporting practices were developed for the new Title I system. If carefully followed, these practices should eliminate many of the sources of error and confusion which have existed up to the present time. Two problems, however, remain: possible bias due to tendencies to report only favorable results, and inadequacies existing in the instruments used to measure gain.

There will probably always be some suppression of negative evaluation findings. There will be suppression motivated by fear that negative results will lead to reduction or termination of funding, suppression motivated by face saving, and suppression motivated by a sincere belief that something must be wrong with the data. The first and third of these types of suppression can be minimized by appropriate assurances--assurances that

funding is not dependent on success and assurances that, because of measurement error, it is indeed possible to obtain negative results with small samples of children who actually benefitted from the instructional treatment.

It seems important that all levels of the evaluation and reporting hierarchy do whatever can be done to minimize the suppression of negative results. At the same time, it seems clear that such efforts will never be totally successful. Consequently, it should be expected that all aggregations of evaluation data will produce systematically inflated results. They will not provide indices which are representative of the impact of Title I at either the Federal or State levels. It may, however, be possible to estimate the extent of the bias.

With respect to instrumentation, it is not quite so clear how one should proceed. Where standardized tests are used, they can be chosen to be as relevant as possible to what the students are being taught. Criterion- or objectives-referenced tests appear to offer a solution to the insensitivity problem--but at the expense of norm-referencing. Although normative data can be developed for any instrument, it is clear that, as curricula and tests become increasingly diverse and specific, the utility of the norms diminishes. For example, if a student studies phonics all year and jumps from the twentieth to the fiftieth percentile on a phonics test, the gain is difficult to interpret. Most students in the norming sample will not have spent a full year on phonics, and it certainly cannot be concluded that the student's other reading skills will also have been raised to the fiftieth percentile. In short, we have no good way to interpret large, norm-referenced gains in isolated skill areas.

It is, of course, not necessary to derive growth expectations from normative data. True or quasi-experimental designs provide alternatives of substantially greater scientific merit. Unfortunately, such designs are often unfeasible in real-world Title I settings.

We have no ideal solution to offer to the instrumentation problem. We feel, however, that it is currently the weakest link in the educational evaluation chain. Until it is strengthened, the information available for making policy decisions will continue to be of marginal utility.

REFERENCES

Bianchini, J. C. Achievement tests and differentiated norms. Paper presented at the U.S. Office of Education Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Washington, D.C., May 1976.

Horst, D. P., Tallmadge, G. K., & Wood, C. T. A practical guide to measuring project impact on student achievement. Washington, D.C.: U.S. Government Printing Office, 1975. (017-080-01460) Also published as Measuring achievment gains in educational projects. Los Altos, CA: RMC Research Corporation, October 1974. (RMC Report No. UR-243)

Roberts, A. O. H. Personal communication, February 1977.